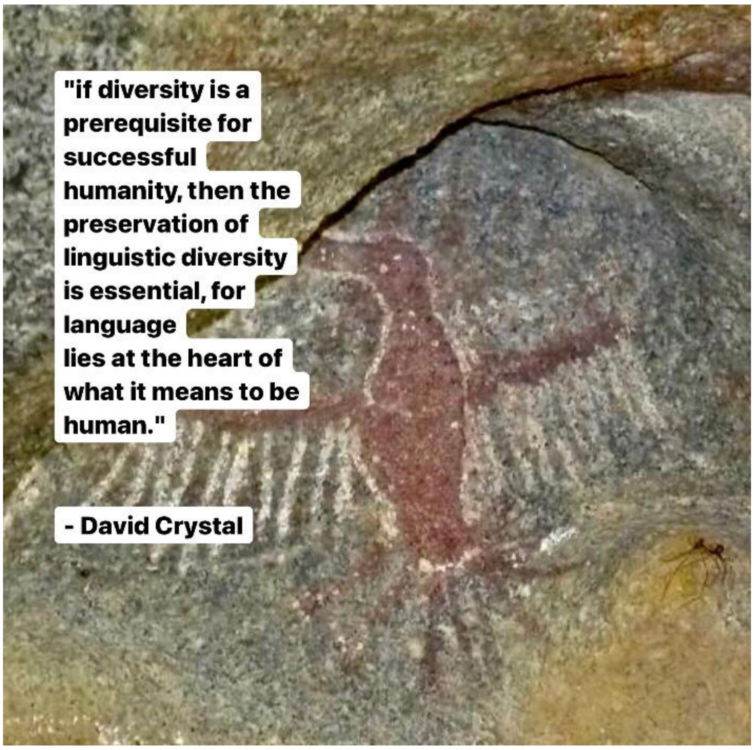
Neural Machine Translation for Papiamento

Preserve culture & identity, unlock new economic value and stay ahead of technological developments.

Kris Croes

12026727
January 2021
University of Amsterdam MBA: Big Data & Business Analytics

In collaboration with:



"if diversity is a prerequisite for successful humanity, then the preservation of linguistic diversity is essential, for language lies at the heart of what it means to be human."

- David Crystal

Disclaimer: *only information cleared for public disclosure is included in this paper.*

Statement of Originality

This document is written by Student Kris Croes, who declares to take full responsibility for

the contents of this document.

I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it. The Faculty of Economics and Business is responsible solely for the supervision of completion of the work, not for the contents.

Acknowledgements

I would like to thank my thesis supervisor Julien Rossi (UVA) for providing guidance during my thesis and also Ali Araabi (UVA) for sharing his deep knowledge on low-resource Neural Machine Translation. Besides the UVA team, I would like to thank Futura/GOA for the opportunity to collaborate on this topic. I am also grateful to Dr. Joyce Pereira, who gladly shared her contacts in the field of Papiamento and who shared her dissertation “Valorization of Papiamento” that served as a great source of inspiration for my thesis. And finally, I want to thank my soon to be wife, Desiree, for her unconditional support during the whole MBA.

Abstract

Government of Aruba (GOA) considers Artificial Intelligence (AI) as a driving force to unlock new (socio)economic value, but government-wide pilots still need to be identified for practical implementation.

Machine translation – part of AI's Natural Language Processing (NLP) domain – has seen major advancements in recent years, with the latest Neural Network-based State-of-the-Art (SOTA) methods (a.k.a. Neural Machine Translation – NMT) lowering the barrier for training low-resource languages.

In collaboration with GOA's Innovation Team and the national innovation Lab (Futura), we investigated what (socio)economic values GOA can unlock with NMT and how these can be unlocked, especially considering Aruba's native language *Papiamento* that is widely used in the Public Domain and which has extremely low digital language resources (e.g. parallel corpus) at the moment.

Based on a qualitative assessment, we identified both societal (e.g. language and identity preservation) and economic benefits (e.g. effort reduction). Furthermore, we trained a first *English to Papiamento* (low-resource) Transformer-based NMT model, which provided measurable insights that helped us in tailoring our advice on specific SOTA methods (e.g. Multilingual training) that could be used for practical implementation of Papiamento-based language models by GOA.

We believe that insights provided in this paper can help GOA with initial directions on implementing practical AI use cases in the NMT domain. Moreover, this initial research together with recommended actions, are assumed to be good building blocks to kickstart the journey of Papiamento towards becoming a higher-resource language, and in doing so catching-up with technological developments.

The source code and datasets created during our research were made publicly available via GitHub to foster further collaboration from the NLP community.

Table of Contents

1. Introduction

1.1. Background

Artificial Intelligence (AI) - a deeply technical family of cognitive technologies, that includes i.a. computer vision, Natural Language Processing (NLP) and robotics, is experiencing one of its peak hypes in the Public Domain due to expected socio-economic benefits [1]. Government of Aruba (GOA) also defined AI as a driving force to unlock new (socio)economic value (i.a. Efficiency, Cost Reduction, Effectiveness and Inclusion/Accessibility) in its recently publish e-Government road map, where government-wide pilots still need to be identified for practical implementation [2].

Machine Translation – the use of computers to automate translation from one language to another – is regarded as one of the biggest advancements in NLP with the introduction of new State-of-the-Art (SOTA) translation methods based on Neural Networks, also referred to as Neural Machine Translation (NMT). While in the past translation systems were expensive and required languages to have millions of sentence pairs (a.k.a. high-resource) available for model development, the latest open-source SOTA NMT methods focused on low-resource languages are lowering both the cost and technical barriers for an estimated 98% of 6.000 languages in the world that lack datasets or machine-readable dictionaries [3].To catch up with technological developments, and in some cases to preserve languages endangered by fast-paced technological developments, practical implementations of NMT are actively being researched and applied for lower-resourced languages [4, 5]. Furthermore, the availability of GPU-powered machines based on per-usage pricing through Public Cloud Services (such as Amazon Web Services or Microsoft Azure Cloud), besides allowing faster training of NMT models, has drastically lowered the cost barrier for practical implementation.

Aruba has a diverse population where Papiamento, Dutch, English and Spanish are widely spoken – with an estimated 76% of the ~100k residents that speak Papiamento in the household. Papiamento – made an official language in 2003 alongside Dutch - is widely used for communication within GOA departments, but as for most languages in the world it lacks digital language resources.

For this research, in collaboration with GOA's Innovation Team and the national innovation Lab (Futura), we investigate what (socio)economic values GOA can unlock with NMT and how these can be unlocked, especially by looking into SOTA NMT methods that can be applied to the low-resourced Papiamento. Besides a qualitative assessment (e.g. literature review and survey), we aim to obtain some measurable results to support our conclusions by experimenting with training a first NMT model to translate from English to Papiamento based on an extremely small -manually collected- dataset of 3.000 sentence pairs.

With the results, we hope to provide GOA with possible future directions on implementing practical AI use cases in the NMT domain, and in doing so helping GOA to think ahead.

1.2. Research Questions

We aim to answer the following three research questions:

1) What (socio)economic values can GOA unlock through the application of NMT?
2) What actions can be taken by GOA to unlock these (socio)economic values?
3) What SOTA NMT methods can GOA use to pilot NMT use cases for the low-resource Papiamento language?

2. Background GOA

2.1. Definition GOA

Government, Public Domain and Public Organization are used interchangeably. The following definition is assumed for Government [6]:

- Formal public organizations that decide on and organize public administration of different sorts and on all levels.
- Organizations that are part of a parliamentary chain of command, which is steered by a set of formal set of rules in order to ensure compliance with political decisions.
- Organizations that vary in terms of function and structure.
- Organizations that deliver services to citizens, both collectively and individually, either directly or by financing private providers.
- Citizens should not be merely seen as consumers but have certain constitutional rights which have to be ensured through rule of law and a fair distribution of social resources.

Besides Citizens, value creation to Civil Servants involved in day-to-day processes to provide such services to Citizens is also considered.

Based on the above definition, we define the scope of GOA as all the official Ministries and underlying Directorates & Service Organizations, including both the Civil Servants that provide Public Services and Citizens that receive such services. The following eight GOA Ministries currently exist across various functional domains:

- Ministry of General Affairs, Integrity, Energy, Innovation and Government Organization.
- Ministry of Infrastructure and the Environment.
- Ministry of Justice, Safety and Integration.
- Ministry of Transport and Communication.
- Ministry of Education, Science and Sustainable Development.
- Ministry of Tourism, Health and Sport.
- Ministry of Finance, Economy and Culture.
- Minister of Social Affairs and Labor.

Aruba is an autonomous country within the Dutch Kingdom through an own parliamentary system, but is dependent on the Netherlands for matters like foreign policy and defense.

2.2. E-Government Road Map

After 2 years in development, in collaboration with i.a. Estonia e-Governance Academy, Columbia University Capstone Project and numerous departments, GOA's e-Government Road Map - named *Building a Better Citizen Experience*- was published in 2020 [2]. Besides economic value principles (e.g. Administrative Simplification, Effectiveness and Efficiency/Cost Reduction), the road map assumes a human-centered approach where e-Government services should also satisfy social-driven principles (e.g. Inclusiveness through multilingual services).

In this road map, AI and 'Open Data' are defined as driving forces to unlock new economic value and enable a pro-active and anticipatory government by 2030. Besides implementing foundational infrastructure and services (e.g. Data Sharing), the road map hence defines the following actions to support AI initiatives:

- Develop a collaboration framework for open data and AI and define a broad AI strategy.
- Develop an open data policy framework and an open data portal/common.
- Ensure government data, information and documents are machine-readable by default.
- Develop a framework and action plan to leverage big data and AI to reduce corruptive practices.
- Build AI talent capacity (both in public and private sector).
- Identify the first use case for AI to pilot and develop a long-term plan to scale it.

Capstone Research Project, Columbia University New York: *"While the common constraints of time, resources, and money apply, Aruba and its residents have shown us that they are ready to take on the challenge of molding their future….."*

The e-Government road map is further prioritized at the highest level in the recent strategic policy framework of Aruba: Master Plan for Economic Recovery and Innovation; Repositioning Our Sails. An aim of the new strategic policy is to increase resilience and reduce the reliance on the tourism-focused model which exposed its vulnerabilities during the COVID-19 pandemic. This policy framework focuses on three main pillars (see *fig.1*), where data driven decision-making and policymaking supported by digital government services are seen as key enablers [7].

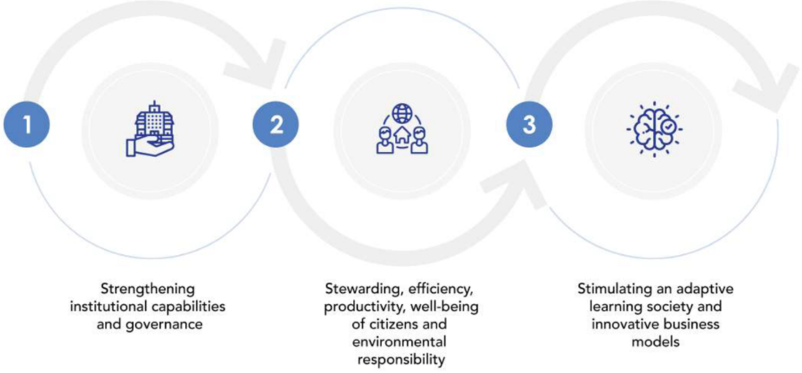

1
2
3
Strengthening institutional capabilities and governance
Stewarding, efficiency, productivity, well-being of citizens and environmental responsibility
Stimulating an adaptive learning society and innovative business models

Figure 1: Main Pillars Aruba Strategic Policy Framework

2.3. Ecosystem Development

To ensure a sustainable and successful roll-out of the e-Government road map, the development of an enabling Ecosystem is further prioritized. The intention is to develop and intensify partnerships with among others other Caribbean nations and Small Island States, Estonia, the Netherlands (and Dutch Kingdom), the European Union and the University of Aruba (in particular the SISSTEM faculty). The latter faculty was established in 2019, in collaboration between the University of Aruba and KU Leuven, focused on offering Bachelor's, Master's and PHD programs on Sustainable Island Solutions through Science, Technology, Engineering and Mathematics while educating a new generation of academics on 21st century skills.

2.4. Futura

Our research was conducted in collaboration with Aruba Futura Foundation (Futura), that resorts under GOA's Chief Innovation Officer. Futura supports GOA in the design, development and implementation of the national innovation strategy including the e-Government road map.

3. Background Papiamento

On the 'ABC' islands Aruba, Bonaire and Curaçao, Papiamento is the most spoken language at home, with Dutch predominantly used as official language in Education and Government [8]. Aruba and Curacao, located near the coast of Venezuela, are each autonomous countries within the Dutch Kingdom.

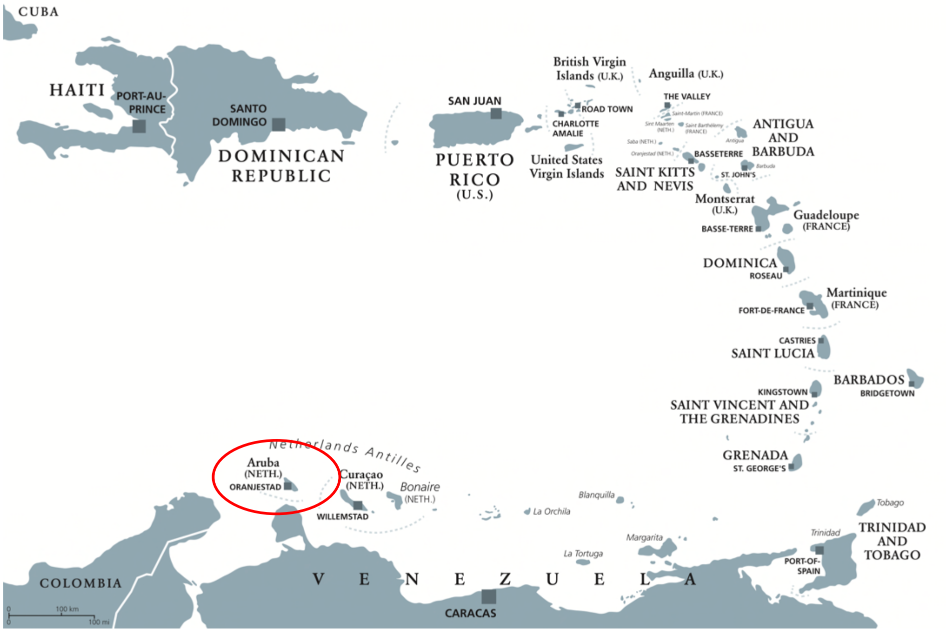


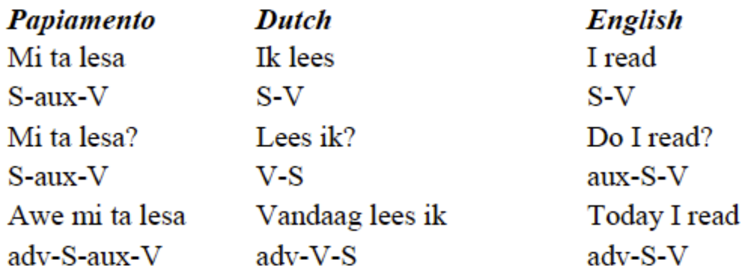
Figure 2: Aruba Geographical Location

Pereira J., 2018 [9]: *"Papiamento is a creole language with a vocabulary of mainly Spanish/Portuguese origin (80 to 85%), to which Dutch, English and to a lesser degree African and native Indigenous (Caquetio) languages also have contributed (Wood, 1972; Maduro, 1953)."*

3.1. Characteristics of Papiamento

Each island has its own spelling and pronunciation. In this research we focus on *Papiamento* spoken in Aruba that has an etymologically oriented spelling compared to the phonologically oriented spelling used for *Papiamentu* in Curacao and Bonaire.

The basic word order in a Papiamento sentence is subject-predicate. Papiamento is classified as a S-(aux)-V-O language, Subject- language, Subject-(auxiliary)-Verb-Object language [10,11]:



Papiamento	*Dutch*	*English*
Mi ta lesa	Ik lees	I read
S-aux-V	S-V	S-V
Mi ta lesa?	Lees ik?	Do I read?
S-aux-V	V-S	aux-S-V
Awe mi ta lesa	Vandaag lees ik	Today I read
adv-S-aux-V	adv-V-S	adv-S-V

Figure 3: SVO examples *Luidens 2015, 2016 [10,11]*

Despite forming to a lesser degree part of Papiamento's vocabulary, there are still many indigenous words that 'survived', mostly related to flora and fauna [12]. One such word is "Blenchi'" ("Hummingbird"), which can be traced back to historical cave paintings made by pre-colonial *Caquetio Indians* settlers.

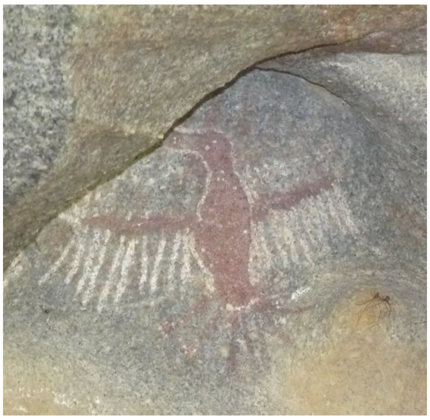


Figure 4: Cave Paintings National Arikok Park Aruba

3.2. Brief History of Papiamento

The development of Papiamento on the 'ABC' islands has its roots in the Dutch Colonial Domination in the 16th and 17th century, where Papiamento became a widely spoken language in these islands.

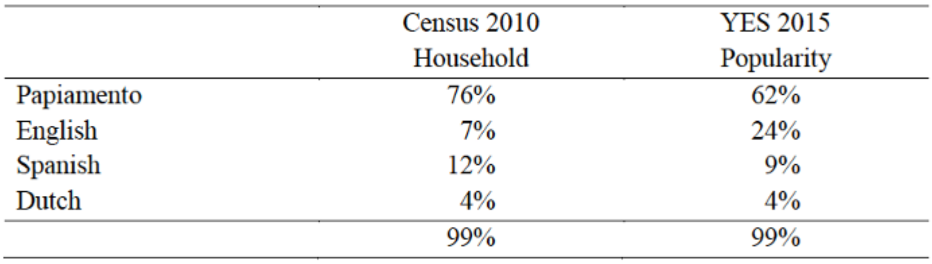
A supported theory for the origin of Papiamento, is that in 15th century an Afro-Portuguese Creole language was already developed and used in commercial transactions (and other social context) between Portugal and African offshore islands of Cape Verde. In the 16th century, in the peak of the slave trade in Curaçao, the language may have been transported from Cape Verde to Curaçao by the enslaved Africans [13]. Here the vocabulary likely further developed within a mixed community of Europeans, Locals, Africans, as well as an indigenous community.

During the Colonial era, Papiamento was regarded as a dialect. Hence, decrees of the Colonial government allowed only Dutch to be used as the medium of instruction in the educational system of the 'ABC' islands long after the Colonial era [9].

3.3. Present day Papiamento

Nowadays Papiamento has a relatively high social profile in the ABC Islands, which resulted in the formulation of an official orthography in 1976. Furthermore, in 2003, the Aruban Parliament enacted the Ordinance on Official Languages, where Papiamento was made an official Language alongside Dutch. This was done with the growing awareness that acknowledging the use of Papiamento on all social terrains as an equally important language was of high importance for the preservation of own culture and identity. This Act on Official Languages stipulates that all official communications within the government can be either in Papiamento or Dutch, both verbally and in writing, and that all documents can be written in both Papiamento and Dutch. Based on this Act, a translation in Papiamento or Dutch can be requested for all documents [9].

Papiamento is now the dominant language spoken by most households. According to the 2010 Census, the total population of Aruba stands at 101,484 persons with 76% of residents speaking Papiamento in the household. Based on a more recent survey conducted by University of Aruba in 2015, Papiamento remains dominant (62%), with English popularity rising among the youth.



	Census 2010 Household	YES 2015 Popularity
Papiamento	76%	62%
English	7%	24%
Spanish	12%	9%
Dutch	4%	4%
	99%	99%

Figure 5: Source: Peterson R., 2015. National Youth Study Aruba, University of Aruba

Papiamento is highly immersed in daily lives and in the social setting. For example, numerous magazines and books are published in Papiamento, music is produced Papiamento (from small local bands to artists being booked on internationally recognized record labels) and there are widespread news and social media sites with daily updated content (e.g. www.bondia.com, www.diarioaruba.com and www.24ora.com).

3.4. Direction of Papiamento in Education Policies

Aruba's official Education policy 2030 bases its foundation on Article 29 of the United Nations Convention on the Rights of the Child. Within this article it is stated that States Parties agree that the education of the child shall be directed to his or her own cultural identity, language and values, for the national values of the country in which the child is living [14].

The majority (72%) of pupils in Aruba speak Papiamento at home but receive Dutch as medium of instruction with equivalent to native Dutch speakers. Research of the Aruba's Department of Education (DEA) shows lower success rates in secondary and higher education for native Papiamento speakers when compared to pupils that speak Dutch at home **[15]**.

After successful pilots, GOA decided to roll-out multilingual education in 2018 across all primary schools with Papiamento as primary medium of instruction, and Dutch and English as foreign languages. The aim is to level the playing field, centered around UNESCO Research showing that a child's ability to scaffold additional language learning (e.g. Dutch in Secondary and Higher Education) is highly increased when encouraging mother tongue development to first achieve cognitive academic language proficiency in a child's native language **[16]**.

To support this new strategic direction, DEA translated several educational materials and books to Papiamento that were originally in English or Dutch, with efforts still ongoing. One such example is the popular Anne Frank's diary that is now also available in Papiamento. To minimize these additional expenses, the use of digital content is prioritized **[15]**. According to *Pereira J., 2018*, a major problem however, is that there is no realistic budget for recruiting qualified staff and producing materials required for educational innovation **[9]**.

Pereira J., 2018 acknowledges the various positive developments around Papiamento, but warns that there is still a risk of endangerment based on the UNESCO scale outlined in the document "Language vitality and endangerment" as it is still excluded as language of instruction in secondary education. The popularization of English among the youth, accelerated by media and television bombardment, is further seen as threat to both Papiamento and Dutch **[9]**.

DEA acknowledges that there is a lot more to be done and considers Aruba to be in a nation-building phase, where innovations in education, in areas such as Papiamento, act as a unifying factor in the society, while keeping an open attitude towards global developments and ensuring a proper transition to academic language proficiency in other foreign languages (such as Dutch and English) **[14]**.

4. Neural Machine Translation (NMT)

4.1. Natural Language Processing (NLP)

Machine translation belongs to AI's domain of Language Technology, also referred to as NLP. NLP is software designed to handle human language in both spoken and written form. NLP is often considered a difficult task in computer science, as the context of language must be learned by a machine to convey the correct meaning depending on the context **[17]**.

NLP can be applied to support one of the following three high-level objectives **[18]**:

1. Aid human-to-human communication (i.e. machine translation).
2. Aid human-to-machine communication (e.g. chatbots, speech or image recognition).
3. Understand language (e.g. meaning extraction from unstructured text).

In the last decade NLP has been growing fast, with the biggest advancement guided by methods based on neural networks (refer to *fig.6*). Neural networks (a.k.a. neural nets) are designed to mimic the function of the neurons in a human brain to ensure better performance. Neural nets are a means of doing machine learning, where a computer learns to perform some task by analyzing training examples (a.k.a. model training).

The concept of neural nets is not new, but high costs of hardware and software (e.g. data storage cost), and slow training of models due to the lack of accessible machines with Graphics Processing Units (GPU) devices, kept the entry barrier high in the past. Today however, Public Cloud Providers are making NLP tasks within reach of most enterprise budgets. Public Cloud Services can be acquired on a pay-per-use basis (e.g. hourly) to support machine learning use cases at a much lower cost compared to buying own hardware and software **[19]**. Furthermore, as part of Public Cloud Services, pre-configured GPU-powered machines for Data Science tasks can be easily acquired to allow faster and cheaper training in the NLP domain (e.g. NMT model training).

Based on the revolutionary potential of AI and NLP, countries and governments are increasing their investments in research and development of NLP technologies to place them ahead of technological developments. Government agencies are applying NLP for example to **[20]**:

- Analyze the feedbacks and comments from internet users (e.g. to understand citizens' concerns).
- Understand citizens' questions. Singapore created for example the chatbot "Ask Jamie" that is trained to virtually assist users on agency websites.
- Enhance policy analysis and facilitate predictions to design measures.

According to a Deloitte report: of the US's Department of Defense total AI spend, NLP has emerged as one of the larger investments with a budget totaling close to US$83 million in 2017, reflecting a jump of nearly 17 percent **[21]**.

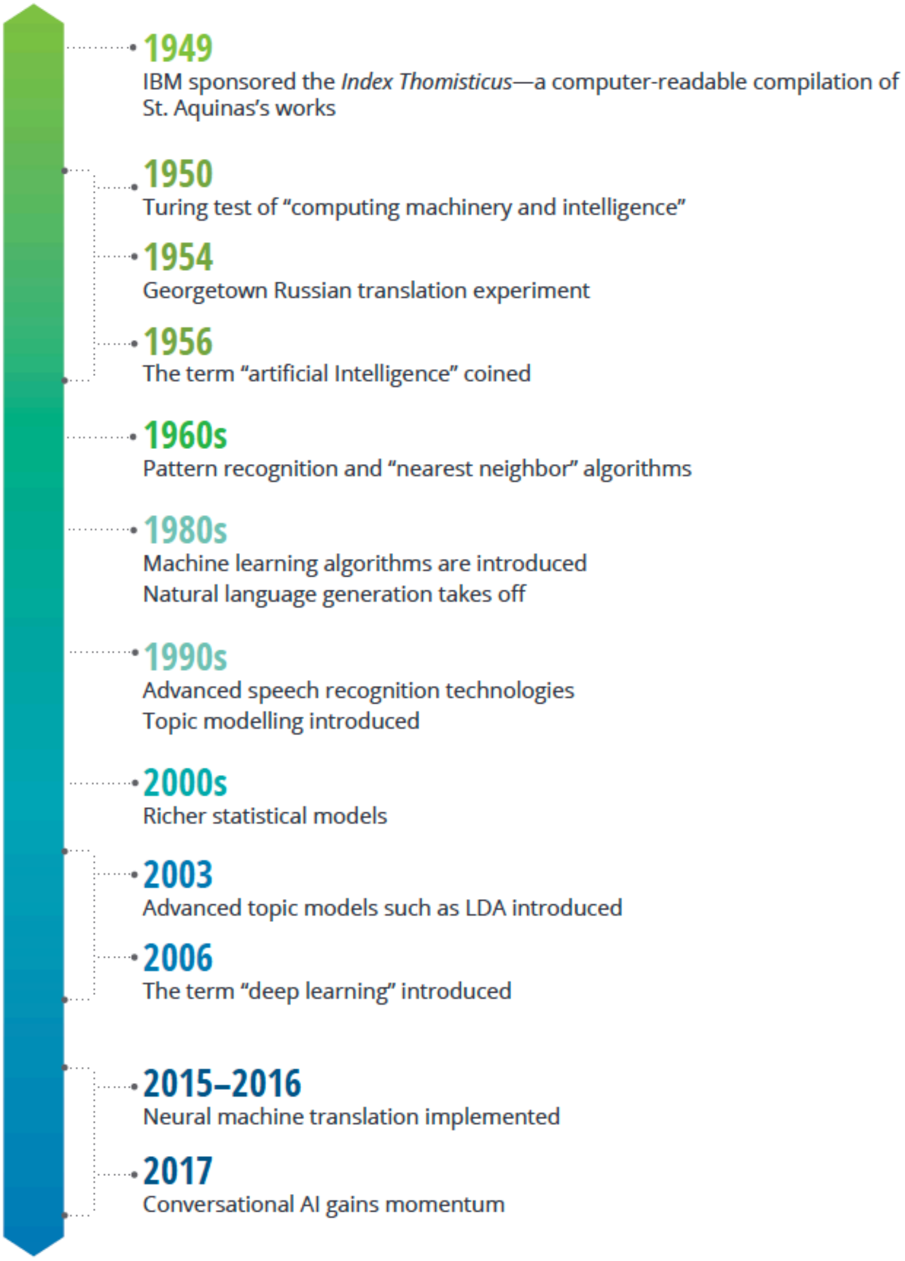

1949
IBM sponsored the *Index Thomisticus*—a computer-readable compilation of St. Aquinas's works
1950
Turing test of "computing machinery and intelligence"
1954
Georgetown Russian translation experiment
1956
The term "artificial Intelligence" coined
1960s
Pattern recognition and "nearest neighbor" algorithms
1980s
Machine learning algorithms are introduced
Natural language generation takes off
1990s
Advanced speech recognition technologies
Topic modelling introduced
2000s
Richer statistical models
2003
Advanced topic models such as LDA introduced
2006
The term "deep learning" introduced
2015–2016
Neural machine translation implemented
2017
Conversational AI gains momentum

Figure 6: Evolution of NLP [20]

4.2. NMT

While past machine translation research and use cases primarily focused on rule-based systems, requiring often extensive manual effort from linguists, since 2016 the focus has shifted to Neural Machine Translation (NMT). Unlike traditional rule-based methods, NMT methods can automatically learn from examples [20].

NMT is the use of computers to automate translation from one language to another, using neural networks, which could be seen as a supervised machine learning task. In supervised learning *“the user provides the algorithm with pairs of inputs and desired outputs, and the algorithm finds a way to produce the desired output given an input. In particular, the algorithm is able to create an output for an input it has never seen before without any help from a human* [22].

To put this in the context of our research, as “teacher” we will provide supervision to the NMT algorithm (i.e. model) in the form of the desired outputs for each input example that they learn from. Specifically, we will provide the model with a list (i.e. dataset) of input sentences paired with the output sentences (i.e. correct sentences translated in Papiamento), in order for the model to learn on itself to translate sentences it has never seen before.

NMT is now a dominant paradigm in academic research with its success mainly attributed to the use of distributed representations of language, enabling end-to-end training of a machine translation system. Unlike classical statistical machine translation systems, separate lossy components like word aligners, translation rule extractors and other feature extractors are not required [23].

4.3. Datasets and Size Implications

To train a translation model, a parallel corpus needs to be collected and fed into the neural network. Parallel corpus means having two lists of parallel sentences. In our research, this implies one list that contains the sentences in Papiamento and the other list that contains translations of the Papiamento sentences to another language (e.g. English), where each line of one file is related to the same line of the other list. Below a sample is shown with three sentences shown for Papiamento to English.

Source sentences	**Target sentences**
Con ta bai?	*How are you?*
Aruba ta parti di Reino Hulandes	*Aruba is part of the Dutch Kingdom*
Nos por uza Technologia di lenguahe	*We can use language technology*

Figure 7: Sample parallel sentences

The amount of training data is seen as a key challenge in the NMT domain. NMT models have a steep learning curve that depend on the dataset size. To get state-of-the art translation accuracy, corpus sizes of millions of sentences are often needed [24]. However, of the estimated 6.000 languages in the world, over 98% lack bilingual corpus or machine-readable dictionaries [3]. This is underpinned by collected data from Google research, showing that only a few languages in the world have large-sized parallel corpus available, ranging from millions to billions sentence pairs. These are defined as high-resource languages (such as English and Spanish). On the lower end, dataset sizes are in the range of 35k, defined as low-resource languages [25]. In some cases, languages are categorized as extreme low-resource, for example when only having 15k or less available parallel data [4,5]. The figure below depicts this distribution.

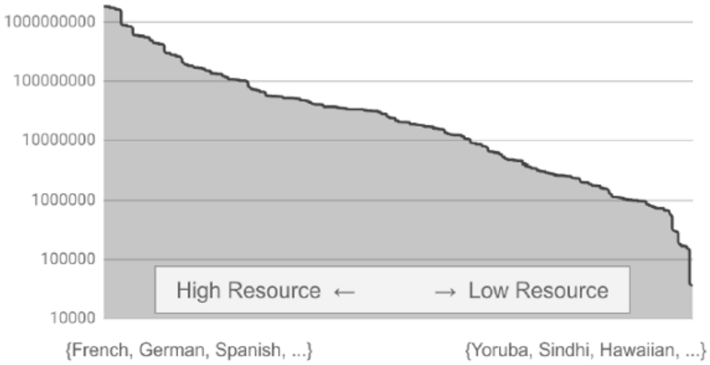


Figure 8: Data distribution over language pairs [25]

Commonly used sources for parallel text include Subtitles (e.g. Ted Talks, Movies, TV shows), parliamentary documents and the Bible. The Bible is often regarded as the most multilingual corpus publicly available, containing 30k sentences in over 900 languages [18, 25]. The dedicated NLP community site http://statmt.org/ is for example a key platform that shares such training data, including evaluations results, to foster collaboration and enable benchmarking capability within the NMT community. Other common approaches to obtain a parallel corpus are 1) scrape the Web via an automated tool, 2) via crowdsourcing 3) or from vendors [26].

In the context of our research, Papiamento could be characterized as an extremely low-resource problem, given limited -readily available- sources of parallel corpora. For example, official documents within GOA (e.g. such as parliamentary proceedings) are still only available in Dutch despite mandated since the officialization of Papiamento. Furthermore, some articles in this officialization Act restrict the use of Papiamento as official language for legislation, judiciary and notarial acts. The latter remains ‘Dutch Only’ [9].

4.4. Preprocessing

For NMT use cases, ‘garbage in, garbage out’ applies, meaning that low quality corpus (a.k.a. noisy datasets) can worsen quality of developed translation systems. This is especially the case for data collected digitally, that tends to be partly noisy (e.g. unstructured text via websites, forums etc.). Hence, there are several advanced cleaning tactics focussed on removing noisy sentences, such as duplicate sentences, sentences with many alphanumerical characters or same sentences that are translated in different ways [27]. For an extremely-low dataset size, these tactics could be counterproductive as the training dataset size might be even more reduced after pre-processing which can further decrease translation quality.

Another common pre-processing step is subword segmentation which can improve the performance of NMT especially for ‘rare words’, that are words that were never seen during training. In the context of our research, the intuition is that subword segmentation could also be useful for Papiamento given some words in Papiamento can be translated via subunits. For example, plurals in Papiamento resemble morphological transformations of West-African languages, where in Papiamento’s case ‘nan’ is added to a word if the noun has a definite article or a deictic pronoun (e.g. Papiamento word for ‘buki’ -meaning ‘book’-, is morphologically transformed to the plural ‘bukinan’ in such cases) [9].

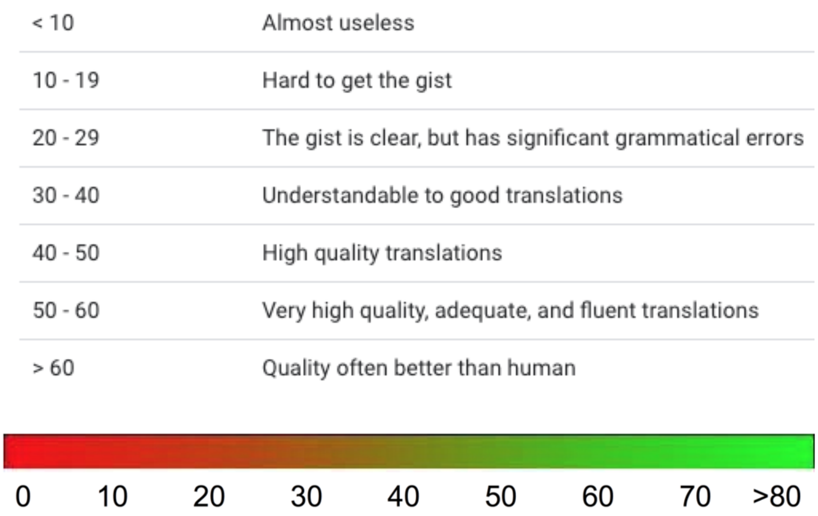
4.5. Evaluation

To determine the accuracy of machine-translated text, Human evaluation is still regarded as the Gold Standard. Humans are more accurate than machines when evaluating metrics such as Fluency, Adequacy and Overlap. However, this can be time-consuming and expensive. Therefore, automated evaluation methods are preferred within NMT, especially if there are a vast number of sentences to be evaluated [18].

BLEU (Bilingual Evaluation Understudy) is the most common used method for automatically evaluating machine-translated text. BLEU, that focusses on assessing overlap, compares produced translations by the model with one or more acceptable reference translations where it looks at the presence or absence of particular words and ordering of words. Study suggests that BLEU scores correlate well with human evaluation of translation quality [28].

The BLEU method produces a number between 0 and 1 (often depicted in percentage), with 0 meaning that the machine-translated output has no overlap with the reference translation (low quality) and value 1 meaning there is a perfect overlap with the reference translations (high quality).

There are no official BLEU scale definitions, however Google suggests the following scale as a rough guideline [29]. For example, based on this scale, a score lower than 20 suggests hard to understand translations, a score between 20 and 40 suggests translations are becoming more understandable and a score above 40 suggests translations are becoming of high quality.



< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

0
10
20
30
40
50
60
70
>80

Figure 9: BLEU scale rough interpretation [29]

Based on a recent study of translated text with Google Translate for example, large variations are seen in BLEU depending on a language being low or high resource. For example, English to Spanish reports an 84 BLEU score while English to Thai and English to Japanese show a BLEU score of only 8 [30].

BLEU scores are also impacted by model settings. For example, based on recent papers, BLEU scores for lower resourced languages can vary between 3.3 and 18. For example, a low-resource Cherokee to English NMT model produced BLEU scores between 6.5 and 15.8 with a 14K corpus size [4], while a German to English NMT model with a more limited corpus size of 10K managed to produce BLEU scores from 3.3 up to 18.3 through hyper-parameter settings [31].

Besides the dataset size and model settings, NMT models seem to perform better when trained on domain specific datasets (a.k.a. in-domain) compared to non-domain specific (a.k.a. out-of-domain) datasets. An example is the Legal domain that has domain-specific topics, wordings and formalities [24].

Another aspect that makes it difficult to train a translation model is the morphological complexity of a language, where a more complex morphology can lead to lower BLEU scores [18], for example for the Cherokee-based model [4].

Despite being widely used, several limitations are observed in BLEU. One limitation is that BLEU penalizes synonyms by looking at exact matches, where for example the score for the word ‘bike’ produced by the model will not be counted if the reference translation has the word ‘bicycle’, despite conveying the same meaning. There are other proposed methods that better take into account semantics (meaning), such as MEANT and BERTscore [18]. Another recent paper even suggests to retire BLEU as the de facto standard metric, and instead to use other metrics such as CHRF, YISI-1 or ESIM in its place which are found more powerful in assessing empirical improvements [32].

5. State-of-the-Art (SOTA) NMT

5.1. Encoder-decoder structure: RNN versus Transformer

SOTA NMT systems are based on an encoder-decoder structure with an attention mechanism, suitable for sequence-to-sequence task like machine translation [33]. Two popular variants are the recurrent NMT architecture (RNN) and the Transformer architecture [34]. In both approaches, the objective of the encoder is to map a source sentence into a sequence of state vectors, whereas the decoder uses the previous decoder states, its last output, and the attention model state to infer the next target word. The attention mechanism (between the Encoder and Decoder) selects and combines the encoder states that are mostly relevant to infer the next word (refer to *fig.10*).

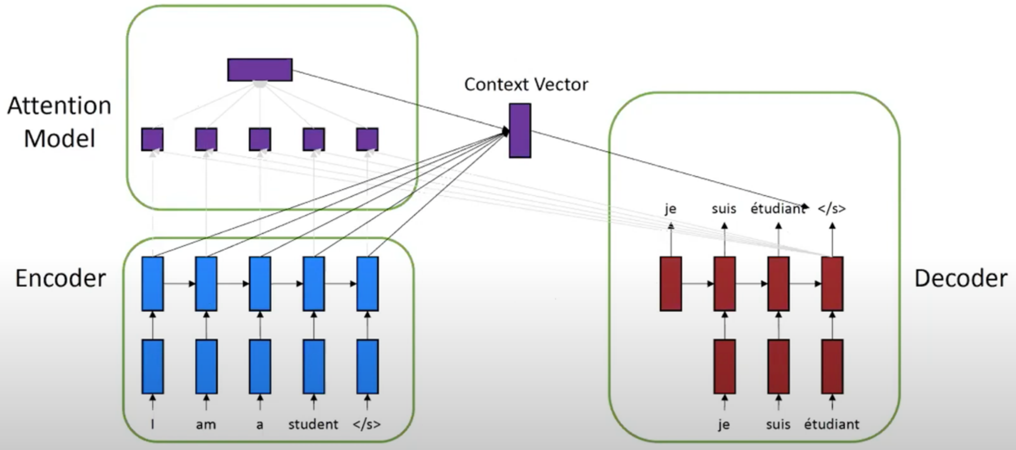


Figure 10: Encode-Decoder model with attention [18,33]

The Encoder and Decoder can be seen as separate neural networks with each a specific objective. For a Machine translation task, you feed the Encoder with collected parallel source and target sentences, that are converted into machine understandable vectors (i.e. sequence of numbers). You then pass this information to the Decoder which tries to learn how to translate source to target based on the encoded information.

Unlike Transformer, RNN has difficulties in learning as many steps may be required given RNN reads one word at a time. In addition, given the sequential nature, it is difficult to fully take advantage of modern multiple GPU computing devices to accelerate training [35]. Based on recent research, for example for English to Chinese translations, Transformers produce better translations compared to RNN [36].

Hence, Transformer has become the de-facto mainstream architecture. The Transformer architecture has achieved great success in NMT and has already been extended to other tasks such as Parsing, Speech Recognition, Speech Translation and Chatbots [35].

For our research, we try to develop a first Papiamento-based NMT system in Chapter 6 using the Transformer architecture.

5.2. Transformer Architecture

The Transformer is the first NMT model that relies entirely on self-attention to compute representations of its input and output without using recurrent neural networks (RNN). The following diagram from *Tubay et al, 2018*, shows a simplified view of Vaswani's **[34]** Transformer architecture.

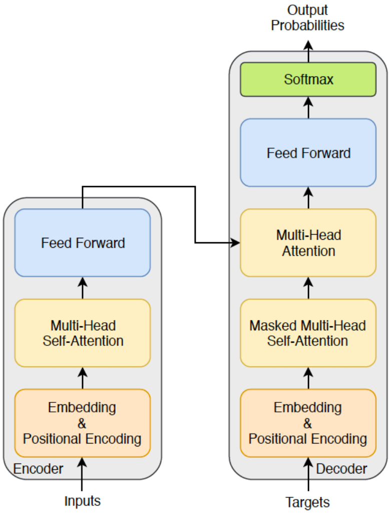

Output Probabilities
Softmax
Feed Forward
Multi-Head Attention
Masked Multi-Head Self-Attention
Embedding & Positional Encoding
Decoder
Targets
Feed Forward
Multi-Head Self-Attention
Embedding & Positional Encoding
Encoder
Inputs

Figure 11: Simplified diagram of the Transformer model

Tubay et al, 2018 **[35]**:

"***The encoder*** *is composed of three stages. In the first stage input words are projected into an embedded vector space. In order to capture the notion of token position within the sequence, a positional encoding is added to the embedded input vectors. The second stage is a multi-head self-attention. Instead of computing a single attention, this stage computes multiple attention blocks over the source, concatenates them and projects them linearly back onto a space with the initial dimensionality. The individual attention blocks compute the scaled dot-product attention with different linear projections. Finally, a position-wise fully connected feed-forward network is used."*

"The decoder *operates similarly, but generates one word at a time, from left to right. It is composed of five stages. The first two are similar to the encoder: embedding and positional encoding and a masked multi-head self-attention, which unlike in the encoder, forces to attend only to past words. The third stage is a multi-head attention that not only attends to these past words, but also to the final representations generated by the encoder. The fourth stage is another position-wise feed-forward network. Finally, a softmax layer allows to map target word scores into target word probabilities."*

5.3. SOTA NMT methods for Low-Resource Languages

Besides applying hyper-parameters to low-resource model training as described in paragraph 4.5., there are also other advanced techniques being researched in order to increase the accuracy of translation of low-resource languages.

A popular approach for example is Google's multilingual NMT architecture introduced in 2017 by *Johnson M. et al* **[37]**. In a multilingual NMT model, besides the small training dataset of the low-resource language, an additional higher-resourced dataset -preferably with some genealogical relatedness- is added to the mix and trained within a single model. Hence, all parameters are implicitly shared by all the language pairs being modelled which forces the model to generalize across language boundaries during training. This approach requires no change to existing SOTA Architectures (such as RNN or Transformer).

Johnson M. et al 2017, identified three multilingual model variants i.e. many-to-one, one-to-many and many-to-many. An example of a *one-to-many* model is training the low-resource English to Catalan (EN→Catalan) together with the high-resource English to Spanish (EN→Spanish), where it is expected for the model to produce better Catalan translations in contrast to training the model solely on a low-resource EN→Catalan corpus. This is because multilingual NMT allows languages belonging to the same genealogical language family (e.g. Catalan and Spanish) to share characteristics during training.

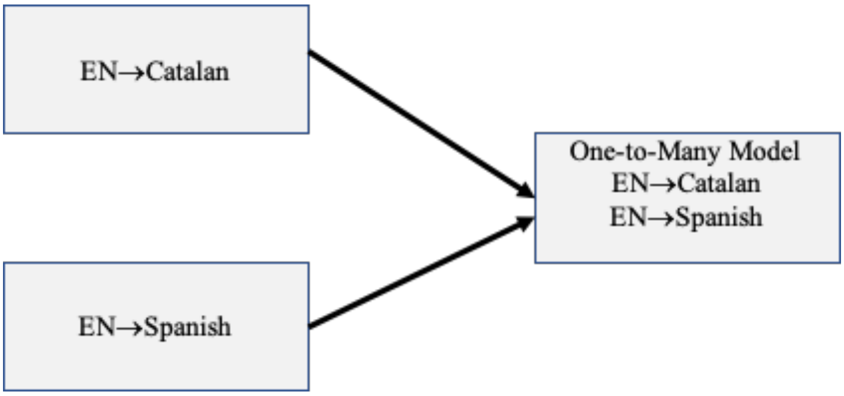


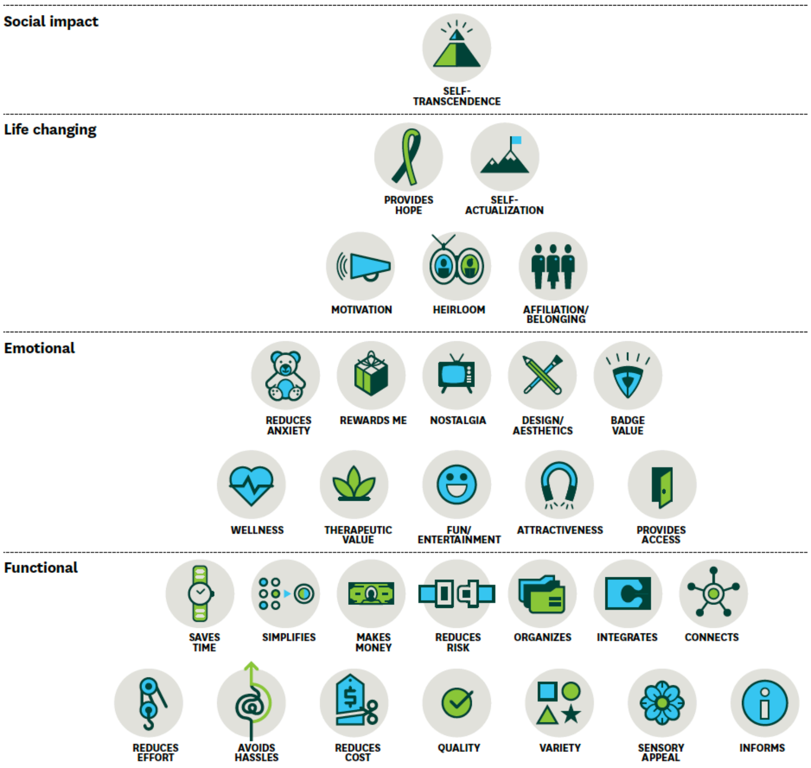
Figure 12: Example One-to-Many multilingual NMT

There are also efforts by researchers to develop a single translation model based on neural networks that try to train a single model to translate to thousands of languages currently not supported (a.k.a. Universal Translation, Massively Multilingual or Massive MT). In this case, a massive amount of different language pairs is added into the same mix for training. For example, there are about 30 researchers at Google working on this problem for the past two years. However, it is still not certain whether this problem can be solved **[18]**.

6. (Socio)Economic Value of NMT

6.1. Definition of Value

To ensure a systematic way to identify and categorize value of machine translation, we apply the 'elements of value pyramid' created by Almquist E. et al. based on three decades of consumer research experience at Bain & Company. This framework extends Abraham Maslow's "hierarchy of needs" and focuses on describing consumer behavior in relation to products and services [38]. According to this framework, products and services deliver fundamental elements of value (a.k.a. value categories) that address four kinds of needs: Functional, Emotional Life Changing and Social impact.



© COPYRIGHT 2015 BAIN & COMPANY INC.

Figure 13: Value Pyramid Scales and Categories

In general, the more categories are covered, the higher the value to Citizens or Civil Servants is expected to be considering the current context. Furthermore, the most powerful forms of perceived value reside on the top in the higher-order elements. The relevance of categories varies according to industry, culture, and demographics. Refer to https://media.bain.com/elements-of-value/ for detailed descriptions per category.

6.2. NMT (socio)economic values unlocked by governments worldwide

Based on literature review of NMT use cases in the Public Domain, both (less tangible) higher-order value categories (i.e. *Heirloom, Belonging and Social Impact*) and (more tangible) lower-order value categories (i.e. *Reduces Cost, Saves Time, Reduces Effort, Provides Access and Reduces Risk*) were identified.

Heirloom, Belonging

Of the estimated 6000 languages spoken in the world, only a few hundred languages have genuinely been given a place in education systems and the Public Domain, and less than a hundred are used in the digital world [4]. There is the chance of language endangerment if speakers cannot make use of electronic technology [39]. Over 98% of world languages lack digital resources to enable machine translation, for example lacking bilingual corpora or machine-readable dictionaries. Digital language resources can help prevent the disappearance of languages and can ensure the development of our tangible and intangible heritage [4].

To revitalize the Cherokee Language for example, for which limited fluent first language speakers remain, NMT is considered in assisting with translations to increase the Cherokee language content that the younger generation is exposed to beyond school hours. Here NMT is used as a method to help spread and preserve Cherokee history and culture [4]. The produced model code and related datasets in the latter paper were made available to the public to attract the interest, and foster collaboration, from the NMT research community with respect to the Cherokee language.

Social Impact

Besides identity and culture preservation, NLP is expected to have large impact on society as a whole given it is expected to boost the GDP of countries in the coming years according to a recent McKinsey paper [40].

Reduces Cost

European Union (EU) sees machine translation as a cost-effective method for making public services available in multiple languages. An example given is the case where a Portuguese entrepreneur is interested in the commercial law of Latvia or Slovenia. It is very unlikely that those documents will be available in Portuguese and manual translation of such documents is seen as more costly than machine translation [41].

Saves Time and Reduces Effort

The Department of Arts and Culture in South Africa does not see machine translation as a replacement of linguists or qualified translators, but rather as a useful tool to aid them with their large translation workloads. Machine translation is used to automatically generate a first draft of translations which the human translator can correct. This speeds up their work in the Public Domain, and in doing so ensuring that as much of the official documentation as possible is available for all South Africans in the language of their choice [42].

Provides Access

EU further sees MT as the ultimate language technology that is able to crack language barrier and play a key role in e-Government services. The idea is that EU citizens should be able to understand the laws or access to public services in their mother tongue [41]. The UN is building for example a considerable archive of parallel documents to support machine translation efforts focused on providing access to published materials in 6 officially mandated languages [43].

Reduces Risk

Several use cases were identified focused on minimizing reliance risks with respect to scarce technological and human resources.

Readily NLP tools often do not support domain-specific machine translation use cases or less commonly used languages (e.g. low-resource languages). Out-of-the-box technological resources are predominantly English-centric, as the industry often prioritizes NLP products that create revenue streams, or in other words: *the industry often prioritizes higher-resource languages due to a profit-driven mindset*. To reduce such reliance risks [44,45]:

- The US government invests in own machine translation applications to support their military operations abroad, as they expect their specific requirements to remain unmet by the commercial sector which supports only a dozen languages [46].
- EFNIL (European Federation of National Institutions for Language) ran a pilot aimed at developing language technology support for the production of bilingual dictionaries which are considered by mainstream publishing houses as commercially unviable [41].

With respect to scarce human resources, Latvia and Lithuania for example use machine translation to reduce reliance on a relatively small pool of human translators that can speak the language fluently [47].

Variety

It is believed that experience with MT is able to branch out in a variety of opportunities in NLP or AI. Existing use cases can for example be leveraged for other AI-driven use cases outside of the machine translation domain. According to Pedtke T. 1997, strategic Federal Government investments in planning, coordination and support of machine translation development can contribute to a more competitive posture in the global Information Age and spin-off benefits across society and other NLP or AI domains [48].

6.3. NMT (socio)economic value opportunities for GOA (survey)

To further identify GOA-specific opportunities, we developed a questionnaire in collaboration with Futura to survey 24 GOA departments across the eight ministries thought to be most relevant based on the amount of data processed and relevancy in the context of the e-Government road map. It was decided to address NLP in a broader sense instead of focussing on only NMT, in order to get an idea of the relative importance of NMT compared to other tasks such as Chatbots and Voice Recognition Systems.

The questionnaire consisted of 14 questions. Questions were mostly open-ended, to get surveyees to share openly and to avoid leading them with own thoughts, inspired by design considerations in *Constable's book Talking to humans 2014* **[49]**. Refer to *appendix A* for the designed survey questions.

During the finalization of this thesis, more than half of the survey results were still pending. Even though initial results show that GOA departments see possible untapped opportunities around both NMT and other NLP tasks (e.g. Chatbots and Meaning Extraction), the conclusions were not yet processed as part of this paper due to lacking statistical significance.

7. Transformer NMT Model: English to Papiamento

To give more substance to the research questions answering and include also a technical point of view, we extend our qualitative assessment with the development a first experimental NMT model for translating English to Papiamento (EN→PAP) using the SOTA Transformer NMT architecture.

We chose two variants for this experiment, a bilingual model to be trained solely on an English to Papiamento (EN→PAP) corpus and a one-to-many multilingual model where (EN→PAP) is to be jointly trained with an English to Portuguese (EN→PT) corpus in a single model according to *Johnson M. et al 2017* **[37]**. The latter choice is based on Papiamento's genealogical relatedness with Portuguese that is assumed to increase the translation accuracy through joint learning with the higher-resourced Portuguese.

The two experiments are carried out using FAIRSEQ, a popular open-source sequence-to-sequence toolkit maintained by *Facebook AI research* that allows the training of custom translation models. FAIRSEQ reduces development complexity and time by providing pre-trained model examples (e.g. multilingual examples) that can be more quickly adapted to your specific use in contrast to building from scratch. It further supports training on (multiple) GPU devices to speed up the training of neural networks **[50]**.

7.1. Corpus preparation

There are limited sources of parallel corpora in Papiamento and there is no readily available corpus that can be used. Based on a desk research, Aruba's official government website was identified as a good starting point to collect parallel sentences. This site has an archive of government news articles in Papiamento, English and Dutch going back to 2012. It was opted to manually collect parallel sentences for PAP→EN. An automated tool for collection (i.e. web scraper) was not built to support with this exercise. In this manual exercise, the copied sentences were aligned in a spreadsheet by first-language Papiamento speakers to increase the probability of correct alignments. As this task was time-consuming and due to time constraints, we were only able to collect 3K sentence pairs (*fig.14 shows some source samples*).

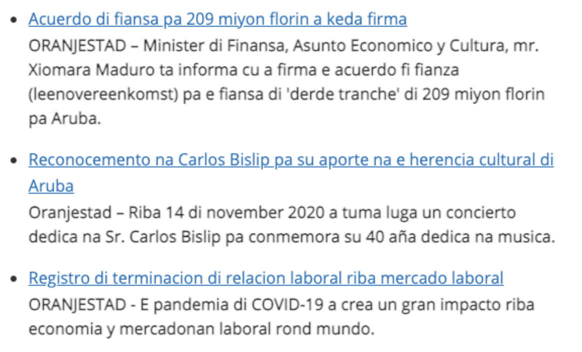
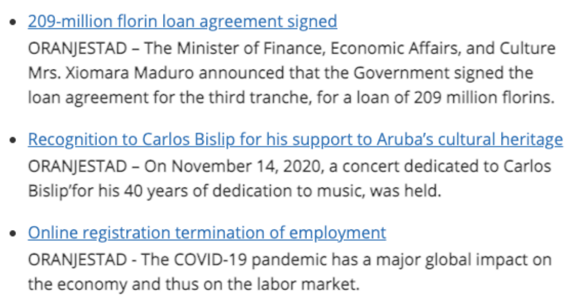
Archive in Papiamento https://www.gobierno.aw/actualidad/noticia_47171/archief	**Archive in English** https://www.government.aw/news/news_47033/archief
 • Acuerdo di fiansa pa 209 miyon florin a keda firma ORANJESTAD – Minister di Finansa, Asunto Economico y Cultura, mr. Xiomara Maduro ta informa cu a firma e acuerdo fi fianza (leenovereenkomst) pa e fiansa di 'derde tranche' di 209 miyon florin pa Aruba. • Reconocemento na Carlos Bislip pa su aporte na e herencia cultural di Aruba Oranjestad – Riba 14 di november 2020 a tuma luga un concierto dedica na Sr. Carlos Bislip pa conmemora su 40 aña dedica na musica. • Registro di terminacion di relacion laboral riba mercado laboral ORANJESTAD - E pandemia di COVID-19 a crea un gran impacto riba economia y mercadonan laboral rond mundo.	 • 209-million florin loan agreement signed ORANJESTAD – The Minister of Finance, Economic Affairs, and Culture Mrs. Xiomara Maduro announced that the Government signed the loan agreement for the third tranche, for a loan of 209 million florins. • Recognition to Carlos Bislip for his support to Aruba's cultural heritage ORANJESTAD – On November 14, 2020, a concert dedicated to Carlos Bislip'for his 40 years of dedication to music, was held. • Online registration termination of employment ORANJESTAD - The COVID-19 pandemic has a major global impact on the economy and thus on the labor market.

Figure 14: Sample 2020 news articles Gobierno.aw

For the multilingual model, the EN→PT dataset was derived from Europe's parliamentary proceedings via https://www.statmt.org/europarl/, a common source used in academia for NMT training and benchmarking.

While seeking parallel corpus in Papiamento, also other potential sources were found, such as the Gospel of Luke that was just recently translated in Papiamento through GOA's Ministry of Education. It is estimated that the current corpus could be amplified 10x times through further collection via remaining government news archives and alternative data sources. For a list of alternative data sources identified during this research, refer to *appendix B*.

7.2. Model Pre-processing

After splitting the data into training, validation and testing sets, we tokenize and pre-process the datasets with the de-facto BPE (Byte Pair Encoding) for subword segmentation. BPE is applied using the open source tokenizer SentencePiece [51], based on a joint vocabulary per experiment. After this, the pre-processed datasets are passed to FAIRSEQ for binarization, training and evaluation.

Despite a reliance on the quality of the original translations, we do not apply other advanced cleaning tactics (e.g. sentence pair removal based on various triggers described in paragraph 4.4.), as we assume this to be counterproductive for our already extremely small dataset. Moreover, common noise was expected to be somewhat reduced through the manual alignment by first-language speakers.

The resulting EN→PAP training dataset consists of 2600 Papiamento sentence pairs. After tokenization and BPE, there are around 6K unique words and 2K unique tokens for both EN and PAP. This suggests that Papiamento does not have a high morphological complexity. The table below shows further statistics of the EN→PAP dataset (*fig.15*).

Language	**Sentence**	**Words**	**Unique Words**	**Tokens**	**Unique Tokens**
pap	2600	50601	5912	79543	2055
en	2600	46371	6468	76342	2038

Figure 15: Sentence, Word and Token statistics EN→PAP

7.3. Model Settings and Results

Extensive exploration of hyper-parameters can be expensive from a computational and cost perspective. Hence, we try to mimic as much as possible recent success booked with hyper-parameters and optimizer settings on low-resource data according to *Araabi A. & Monz C. 2020* [31]. In this paper, higher BLEU scores were achieved through tailored settings per dataset size, in contrast to the Transformer baseline settings introduced in 2017 by *Vaswani A. et al* [34]. We conduct four experiments:

- **Exp1:** EN→PAP transformer model, with training, validation and test dataset sizes of respectively, 2600, 100 and 300, applying primarily default transformer settings.
- **Exp2:** same as Exp1, except for the application of optimal settings found for a 5K dataset size in *Araabi A. & Monz C. 2020.*
- **Exp3:** EN→PAP/ EN→PT multilingual transformer model, with training, validation and test dataset sizes of respectively, 2600, 100 and 300, for both EN→PAP and EN→PT. The same as settings as Exp2 are applied for this case.
- **Exp4:** same as Exp3, except for 10x EN→PT training, validation and test dataset sizes, namely 26000, 1000 and 3000. Here the optimal settings for a 20K dataset size found in *Araabi A. & Monz C. 2020* are applied.

Model parameters were further optimized with the commonly used Adam optimizer **[52]**. As indicated by *Johnson M. et al 2017* **[37],** our main (Transformer) architecture remained the same across the two variants (bilingual<>multilingual), except for some multilingual-specific FAIRSEQ settings and the additional dataset (EN→PT) added to the (EN→PAP model).

The training was done in batches of max 4000 tokens at a time and for max 100 epochs, configured to stop if the model converges before that time through loss-function minimization. After training, the models were evaluated on the test sets and the BLEU scores generated using SacreBleu **[53]**, with BEAM size 6.

To run experiments and to speed up training, a temporary a Data Science Virtual Machine (DSVM) was rented on an hourly basis from Microsoft Azure Cloud. This DSVM had a single Tesla V100 GPU (~3 EUR/hr). Even though training of the bilingual model was doable on a CPU-only laptop (~ 4-6 hours), training the multilingual model on a CPU-only device was found less practical.

The specific settings and results are shown per experiment in the following table (*fig.16*).

	Exp 1 EN→PAP	**Exp 2** En→Pap	**Exp 3** En→Pap En→Pt	**Exp 4** En→Pap En→Pt
FAIRSEQ Arch	Transformer	Transformer	Multilingual transformer	Multilingual transformer
Training size En-Pap/En-Pt	2600	2600	2600/2600	2600/26000
Validation size En-Pap/En-Pt	100	100	100/100	100/1000
Test size En-Pap	300	300	300	300
BPE size	3000	3000	6000	16000
Feed-forward dimension	2048	512	512	1024
Attention heads	8	2	2	2
Dropout	0.1	0.3	0.3	0.3
Enc/dec layers	6	5	5	5

Label smoothing	0.1	0.6	0.6	0.5
Enc/dec layerDrop	0.0/0.0	0.0/0.3	0.0/0.3	0.0/0.2
Activation dropout	0.0	0.3	0.3	0.3
Results				
BLEU score: En->Pap	8.77	12.57	13.68	22.12
Training duration	~ 5 minutes	~ 5 minutes	~ 10 minutes	~ 2 hours

Figure 16: NMT model EN→PT settings and BLEU scores per experiment

7.4. Discussion of Results

By using the SOTA Transformer architectures and applying SOTA best-practices (e.g. BPE, Hyper-parameters and Multilingual model training), a maximum BLEU score of 22.12 was achieved for a model that translates from English to Papiamento based on an extremely low training set ~ 3.000 Papiamento sentence pairs. The score of 22.12 in the multilingual Exp4 is a significant improvement (+ 13.35) when compared to the bilingual baseline model in Exp1 (8.77). This suggests that the translation quality of En→Pap is significantly improved when enriching the training data with a larger-sized genealogical related dataset (in this case an En→Pt dataset 10x the size of the En→Pap dataset), and training En→{Pap, Pt} into a single one-to-many model. Likely, implicit parameters sharing, and generalization across boundaries, occurred during training.

It is noticeable that our highest score of 22.12 even outperforms some recent low-resource benchmarks described in paragraph 4.5., which average around a maximum of 16/18. Besides the multilingual architecture and hyper-parameter settings, the assumption of Papiamento not having a high morphological complexity, could have further contributed to the outperformance in some cases.

The following table (fig.17) shows the translation results per experiment for one of the sentences in the test dataset. An improvement in translation is noticeable per consecutive experiment, although some crucial words are still not properly learned, such as the verb 'organisa' meaning 'organize'. We further asked a couple of first-language Papiamento speakers to rate which experiment produced the best result, with the most answers resulting in either Exp3 or Exp4. This lightly indicates that BLEU is suitable for our Papiamento-based models due to the suggested correlation with human judgement, which is still regarded as the Gold Standard.

EN original sentence	The members of the Love4Art Studio will present their annual exposition at the public library from December 3 till December 21, 2018
PAP ref. translation	Miembronan di Love4Art Studio ta organisa nan exposicion anual na Biblioteca Nacional Aruba di 3 pa 21, 2018.

Exp1	E concepto di december 2019 te cu ta consisti di december 2019 te cu 31 di december 2019 te cu 31 di december 2019 te cu 31 di december 2019 te cu 31.
Exp2	Bronan di e miembronan di Br4 Sr. Brica e exposicion di presenta na biblioteca na Biblioteca na biblioteca di 31 di december 2018.
Exp3	miembronan di e miembronan di Studio aki lo tin un exposicion anual na Biblioteca Nacional na biblioteca di december 2019 na biblioteca december 2019.
Exp4	Miembronan di Ltim'e Stuy exposicion anual na biblioteca nacional di 3 te cu 21 di december

Figure 17: Evaluation results FAIRSEQ for a sentence sample across experiments

Based on Google's BLEU scale definitions described in paragraph 4.5., we managed to develop a model that is becoming more understandable but that still has significant errors, which corresponds with the results in *fig.17*. This suggest that there is still a long way to go to produce human quality translations (e.g. 40 or above). For our research purposes however, this exceeded our expectations. These experiments not only underpin the SOTA techniques that could be used for the GOA/Papiamento-specific context, but also prove that nowadays the development of SOTA NMT models (e.g. also used by leading research giants such as Google and Facebook) can be realized more easily (e.g. via opensource tools that abstract the complexity), at lower budget (e.g. via Cloud Services) and fairly fast (e.g. training via GPU device).

Even though there is still a long way to go to produce human quality translation, these experiments show that even with an extremely small dataset some learning can be done which can lead to good starting points for practical implementation of AI-pilots in the NMT domain. Moreover, our experiments show that there are concrete -achievable- actions that can be taken to enable and accelerate the journey of Papiamento towards becoming a higher-resource language, and in doing so catching-up with technological developments, one being as simple as: *building parallel (corpus) datasets in Papiamento*.

7.5. Availability Dataset and Source code

To encourage further collaboration from the NLP community, the datasets and source code are made publicly available via GitHub: https://github.com/kriscroes/nmtpapiamento2020.

8. Answering the research questions

8.1. Research Question 1

- **What (socio)economic values can GOA unlock through the application of NMT?**

Based on the review of among others global NMT global use cases in the Public Domain and relevant GOA plans and policies, we identify both societal and economic benefits. These are described below.

Preserve Language, Identity and Culture

Papiamento had many positive developments in the past two decades, such as being officialized, having an official orthography released, and now functioning as the primary medium of instruction in primary education. Despite these positive developments, there is still a risk of language endangerment, even though not high compared to other low-resource languages. Indications of endangerment are: 1) Papiamento still needs to be properly embedded in secondary education, 2) Papiamento lacks electronic (language) technologies and 3) English-centric media and television bombardment is increasing the popularity of English among the youth.

NMT can help in preserving the culture and identity for current and future Aruban generations by increasing the availability of electronic (language) technologies that support and understand Papiamento's vocabulary. This can further strengthen Aruba's educational foundations that aim to provide education in a child's own cultural identity, language and values, according to Article 29 of the United Nations.

Reduce Cost, Save Time and Reduce Effort

DEA is tasked with the creation of multilingual material in the education domain and GOA aims to roll-out multilingual e-Government services. As for the usual public entities, they have to work with resource constraints while addressing these objectives. NMT could be used to aid linguists or translators in producing translations at a lower cost and with less effort. A good example, as described in paragraph 6.2., is the Department of Arts and Culture in South Africa that uses machine translation to generate a first draft of translations which the human translator can correct. In such a case, NMT is applied as 'aid to', instead of 'replacement of', qualified personnel, in line with GOA's "human centered" approach.

Further note that the cost of NMT implementation is drastically lower than in the past, through the use of among others open-source tools and usage-based Public Cloud services.

Provide Access

Besides Papiamento, also Spanish, English and Dutch are spoken in Aruba. NMT can be extended to also capture other languages, making e-Government services better accessible to all citizens.

Catch up with Technological developments and Spin-off other benefits

There are no out-of-the-box technological resources for Papiamento. As the commercial sector prioritizes languages that are most used, we do not expect Papiamento to be supported out-of-the-box anytime soon. Although researchers are experimenting with universal machine translations that focus on translating 'forgotten languages' in the commercial sector, it is not certain if this will succeed, when this will succeed, and to what extent Papiamento will be included even if it succeeds on long-term. Proactively kickstarting AI pilots by GOA in the NMT domain can help in catching up with technological developments around digital language resources, and in doing so help "*Papiamento in its journey towards becoming a higher-resource language"*. Furthermore, experience with NMT can branch out in other opportunities in other NLP or AI domains. This in turn can reinforce GOA's objectives, and accelerate results, geared towards unlocking new (socio)economic values with AI.

8.2. Research Question 2

- **What actions can be taken by GOA to unlock these (socio)economic values?**

We propose the following actions, considering the specific context analyzed of GOA/Papiamento.

Build Papiamento Corpus

Our research, and our experiments in chapter 7, show that larger datasets can drastically increase the quality of NMT. The following can be considered:

- Further build on datasets collected in our research. To speed up the collection, an automated scraping tool can be considered which was not built in our experiment.
- Further investigate the usability of alternative data sources encountered during our research, *described in Appendix B*.
- Consider the creation of transcripts in Papiamento for online videos or television programs that are now mostly English-centric. Besides enriching datasets in Papiamento, this can be used to expose the younger generation beyond school hours to content in Papiamento to minimize language endangerment.
- Based on existing legal basis (i.e. officialization act of 2003), mandate parliamentary documents (e.g. proceedings) to be translated also in Papiamento besides Dutch. As reference, it can be looked at how EU is doing this task.
- Extend the Officialization Act to also include Papiamento (alongside Dutch) as official language for legislation, judiciary and notarial acts. This drastically increases data sources that can be used for corpus building.

Prioritize NMT as part of e-Government road map and eventual pilot(s)

A defined action in the e-Government road map is to pilot an AI use case. Considering the concrete NMT (socio)economic values identified for GOA in our research, NMT could be a good candidate for such a pilot (or for future pilots). To increase the probability of pilot success, the results of the -to be completed- survey (described in paragraph 6.3.) can be used to prioritize GOA ministries and/or departments based on for example perceived quick wins or impact. Our literature review also shows that domain-specific use cases are more easily trained, which could be scoped as part an eventual pilot to realize early wins. Moreover, our experiments in chapter 7, including the source code, can serve as baseline and inspiration towards practical NMT implementation by GOA.

Extend e-Government ecosystem development with NMT collaboration

Even though technical and cost barriers for the practical implementation of NMT has become much lower for Public Organizations, skills in this domain will still need to be developed locally by GOA to increase the probability of success. Hence, we advise to partner up with both local academic institutions (e.g. SISSTEM faculty) and academic institutions within the Dutch Kingdom (e.g. Dutch Universities experienced in this domain) to address knowledge sharing and skills development. GOA could also discuss with local faculties the possibilities to include NLP topics in their offerings (e.g. part of local Bachelor's, Master's or PHD offerings).

8.3. Research Question 3

- **What SOTA NMT methods can GOA use to pilot NMT use cases for the low-resource Papiamento language?**

Our technical experiments in chapter 7 show that applying the latest SOTA NMT Transformer Architecture, combined with other novel techniques such as multilingual training and hyper-parameter settings, can produce low-resourced Papiamento-based models that resemble, and in some cases even outperform, recent benchmarks in the low-resource domain. Even though the trained models in our experiments still produce Papiamento translations with significant errors, they can serve as good baselines to further build on towards achieving more understandable translations. It is clear however that larger datasets in Papiamento are required to obtain higher-quality translations. This is because improvements that can be gained with SOTA NMT techniques will reach a limit at some point when lacking sufficient data.

9. Conclusion and Future Work

9.1. Conclusion

Papiamento – nowadays widely used in Aruba’s Public Domain- endured many centuries, even the colonial periods where it was marked as a dialect thought soon to be forgotten. It is remarkable that even some indigenous words still exist today in Papiamento’s vocabulary that were passed on from early indigenous communities.

Even though now an official language, Papiamento has stayed behind of technological developments, currently characterized as very low-resourced due to the absence of digital language resources.

On a positive note, there seem now to be momentum in the Public Domain to accelerate government-wide digitization, where AI is regarded as a driving force to unlock new (socio)economic values. As remarked by Columbia University’s Capstone project, “*While the common constraints of time, resources, and money apply, Aruba and its residents have shown us that they are ready to take on the challenge of molding their future…..”*

In our research, we determined that NMT –part of AI’s NLP domain- could be a good candidate for GOA towards practical implementation of AI. We expect that by including NMT in the e-Government road map and eventual pilots, GOA will be able to unlock both societal benefits (e.g. language and identity preservation) and economic benefits (e.g. reduce effort in creating multilingual content across the Public Domain). Moreover, we suggest that economic benefits can be derived without the need for replacing qualified translators, but rather as an aid to –already constraint– human resources, in conformance with GOA’s human-centred approach.

We also proved with our technical experiments that the barriers to start developing Papiamento-based NMT models have significantly lowered. Through the use of open-source tools (i.e. FAIRSEQ), the renting of a GPU-based cloud machine (i.e. Microsoft Azure) and the application of the latest SOTA NMT techniques (i.e. multilingual Transformer architecture) while using an extremely small dataset (i.e. ~3K), we were able to relatively quickly and cheaply train a first En→Pap model that produces already semi-understandable translations in Papiamento. Our results, including source code, could be used as starting point to further build on, and to support direction on practical pilot implementations.

To achieve these benefits, as described in our analyses, it is important for GOA to be proactive in taking own actions, as digital resources for less-commonly used languages are not expected to be made available out-of-the-box any time soon by the profit-minded commercial sector. Besides fostering the creation of ‘open’ datasets through policies and actions, GOA can create an enabling environment through the inclusion of NMT in concrete future pilots and partnerships (ecosystem development) for skills building.

We believe that GOA is ’not too late in the game’, meaning that taking concrete actions now can help Papiamento in becoming a higher-resourced, and in doing so catching-up with technological developments in NMT, let alone expected spin-off opportunities in other NLP or AI domains.

9.2. Limitations and Future Work

Our NMT model was limited to an extremely small dataset. This remains a challenge when thriving for high BLEU scores. Future work should focus on further amplifying the datasets for Papiamento. Instead of manually collecting the data, a scraper can be considered to automate data collection.

Based on initial -unfinished- survey results, GOA departments see possible untapped opportunities around both NMT and other NLP tasks (e.g. Chatbots and Meaning Extraction). Future work can consider also other NLP tasks determined to be relevant from the finalized survey results.

Our SOTA techniques focussed primarily on supervised NMT techniques. There are also other SOTA low-resource techniques that can be considered in future work, such as Data Augmentation through semi-supervised learning.

References

[1] Kuziemski M. & Misuraca G., 2020. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecommunications Policy. Volume 44, Issue 6, 101976.

[2] Department of Innovation Aruba, 2020. e-Government Roadmap. Ministry of General Affairs Aruba.

[3] Fraisse A. et al, 2019. A Sustainable and Open Access Knowledge Organization Model to Preserve Cultural Heritage and Language Diversity. Information. 10. 303.10.3390/info10100303.

[4] Zhang Z., Frey B. & Bansal M., 2020. Cherokee-English Machine Translation for Endangered Language. ACL, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 577–595.

[5] Gu J. et al, 2018. Universal neural machine translation for extremely low resource languages. arXiv:1802.05368.

[6] Lindgren I. & Jansson G., 2013. Electronic services in the public sector: A conceptual framework. Government Information Quarterly 30, 163–172.

[7] Committee Economic Recovery and Innovation Aruba, 2020. Aruba's Mission Driven Model for Economic. Ministry of General Affairs Aruba.

[8] Delgado S. et al, 2016. Education, Languages in contact, and popular culture in the Hispanophone, Francophone, and Dutch Caribbean. Research Gate.

[9] Pereira J., 2018. Valorization of Papiamento in Aruban society and education, in historical, contemporary and future perspectives. University of Curacao, Research Institute. UCRI Publication No 1.

[10] Luidens M., et al, 2015. Manual di gramatica di Papiamento: sintaxis. Aruba: Departamento di Enseñansa.

[11] Luidens, M. el al, 2016. Preposicion y interheccion di Papiamento. Aruba: Departamento di Enseñansa Aruba.

[12] Buurt G. ,2015. Caquetío Indians on Curaçao during colonial times and Caquetío words in the Papiamentu Language. 10.13140/RG.2.1.3883.6649.

[13] Martinus F., 1996. doctoral dissertation titled, The Kiss of a Slave (1996). Universiteit van Amsterdam.

[14] Coordination Team National Education Plan, 2009. National Education Policy Aruba 2030. DEA.

[15] Department of Education – Policy Department, 2016. Policy Note: Language Diversification in medium of instruction for primary education. DEA.

[16] Ball J., 2010. Educational equity for children from diverse language backgrounds: Mother tongue-based bilingual or multilingual education in the early years. Presentation to UNESCO International Symposium: Translation and Cultural Mediation.

[17] Jurafsky D. and Martin J., 2009. Speech and Language Processing (second edition). Pearson Hall.

[18] Neubig G., 2020. Lecture Slides NLP Low-resource Bootcamp Carnegie Mellon University (CMU). Retrieved from https://github.com/neubig/lowresource-nlp-bootcamp-2020.

[19] Bhattacharya S. & Patil A., 2020. Gain deeper insights and fuel innovation with cloud AI/ML. Retrieved from https://www2.deloitte.com/us/en/pages/consulting/articles/cloud-and-machine-learning-they-are-better-together.html.

[20] Zhang D., Wang J. & Sun M., 2020. The Progress That Natural Language Processing Has Made Towards Human-level AI. Journal of Artificial Intelligence Practice, Vol. 3: 38-47.

[21] Eggers W., 2019. Using AI to unleash the power of unstructured government data. Retrieved from https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html.

[22] Muller A. & Guido S., 2017, Introduction to Machine Learning with Python: a Guide for Data Scientists. O'Reilly Media.

[23] Dabre R., 2020. A Comprehensive Survey of Multilingual Neural Machine Translation. arXiv preprint arXiv:2001.01115.

[24] Koehn P. and Knowles R., 2017. Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39.

[25] Arivazhagan N. et al, 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. arXiv:1907.05019. Google AI.

[26] Tracey J. et al, 2019. Corpus building for low resource languages in the DARPA LORELEI program. In Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages, pages 48–55.

[27] Rikters M., 2018. Impact of Corpora Quality on Neural Machine Translation. arXiv e-prints.

[28] Kishore P. et al, 2020. BLEU: A Method for Automatic Evaluation of Machine Translation. ACL.

[29] Google Cloud Translation Documentation: Evaluating models. Retrieved from https://cloud.google.com/translate/automl/docs/evaluate.

[30] Aiken, M., 2019. An updated evaluation of Google Translate accuracy. In Studies in linguistics and literature, 253–260.

[31] Araabi A. & Monz .C, 2020. Optimizing Transformer for Low-Resource Neural Machine Translation. arXiv: 2011.02266.

[32] Mathur N. et al, 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4984–4997, Online. Association for Computational Linguistics.

[33] Bahdanau D. et al, 2016. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473.

[34] Vaswani A. et al, 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pp. 5998–6008.

[35] Tubay B. & Ruiz Costa-Jussà M., 2018. Neural Machine Translation with the Transformer and Multi-Source Romance Languages for the Biomedical WMT 2018 Task. Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 667–670.

[36] Yuying Y. & Toral A., 2020. Fine-grained Human Evaluation of Transformer and Recurrent Approaches to Neural Machine Translation for English-to-Chinese. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation.

[37] Johnson M. et al, 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

[38] Almquist E., Senior J. & Bloch N., 2016. The Elements of Value. Harvard Business Review, September.

[39] Crystal D., 2002. Language Death. Cambridge University Press.

[40] Bughin J. et all, 2018, Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey discussion paper.

[41] Rivera R. et al., 2017. Language equality in the digital age: Towards a Human Language Project. European Parliamentary Research Service Scientific Foresight Unit (STOA), PE 581.621.

[42] McKellar, C.A., 2014. An English to Xitsonga statistical machine translation system for the government domain.

[43] Ziemski M., 2016. The United Nations Parallel Corpus v1.0. European Language Resources Association (ELRA), LREC'16.

[44] Anastasopoulos A., 2020. Should All Cross-Lingual Embeddings Speak English?. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

[45] Dahlmeier D., 2017. On the Challenges of Translating NLP Research into Commercial Products. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.

[46] Jones D. et al., 2009. Machine Translation for Government Applications. Lincoln Laboratory Journal, Volume 18, Number 1.

[47] Vasiljevs A. et al, 2014. Machine translation for e-Government – the Baltic case. Proceedings of AMTA 2014, vol. 2.

[48] Pedtke T., 1997. U.S. Government support and use of machine translation: current status. Proceedings of MT Summit, San Diego, pp. 3-13.

[49] Constable G., 2014. Talking to Humans 1st ed., Retrieved from: http://www.talkingtohumans.com/.

[50] Ott M. et al, 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

[51] Kudo T. & Richardson J., 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.

[52] Kingma D.et al, 2015. Adam: A Method for Stochastic Optimization. In International Conference on Learning Representations (ICLR).

[53] Post M., 2018. A call for clarity in reporting BLEU scores. arXiv preprint arXiv:1804.08771v1.

Appendix A: GOA Survey Design

1. **Approximately what proportion of data needed to run your department is based on unstructured text data?**

   □ 75% Structured Data / 25% Unstructured Data
   □ 50% Structured Data / 50% Unstructured Data
   □ 25 % Structured Data / 75% Unstructured Data

2. **Which processes create the most unstructured text data or are mostly supported by unstructured text data (select all that apply).**
   □ Internal Operations
   □ Customer Facing (e.g. services to customers)
   □ Marketing/Communication
   □ Other

3. **How do you view the importance of unstructured text data in achieving department goals?**

4. **What are your main challenges in using unstructured text data?**

5. **Is text generated within your department subject to domain specific vocabulary? If applicable, how important is it for your department to adhere to certain domain specific language or vocabulary?**

6. **Is there any organizational or legal mandate applicable for your department concerning text translation? For example, is it mandatory to translate certain documents to/from Papiamento, Dutch or English?**

7. **Are any of the following language technologies already used within your department?**
   1. Machine Translation of Text
   2. Dialogue Systems /Chatbots
   3. Meaning Extraction from Text
   4. Speech recognition
   5. Image recognition

8. **If the above described language technologies are not yet used within your department, were these ever considered or are these being considered at the moment? If so, which of the use cases and in what way?**

9. **Do you see any opportunities to applying one or more of the below language technologies? If so, which processes would benefit the most (e.g. internal, customer-facing etc.) and in what way (e.g. Quality Increase, Efficiency Gains or Cost Reduction)?**

   Machine Translation of Text. Please elaborate.
   Dialogue Systems /Chatbots. Please elaborate.
   Meaning Extraction from Text. Please elaborate.
   Speech Recognition. Please elaborate.
   Image Recognition. Please elaborate.

10. **Do you see any challenges to applying above language technologies (e.g. domain specific vocabulary, readiness for change, IT capability etc.)?**

11. **Which of the following use cases do you expect to have the most untapped opportunities for your department? (choose only one)**
    □ Machine Translation of Text
    □ Dialogue Systems /Chatbots
    □ Meaning Extraction from Text
    □ Speech recognition
    □ Image recognition
    □ None of the above

12. **Is there any other thing that you want to add that we forgot to ask you with regards to language technologies (can be anything)? (optional).**

13. **How 'Ready' do you think the following sectors in Aruba Government are for Language Technologies as part of eGovernment. 1 means not 'ready at all' and 5 means 'fully ready/ripe for change'. (Select only one per sector)**
    Education
    □1 □2 □3 □4 □ 5
    Infrastructure
    □1 □2 □3 □4 □ 5
    Justice / Integrity
    □1 □2 □3 □4 □ 5
    Transportation
    □1 □2 □3 □4 □ 5
    Social
    □1 □2 □3 □4 □ 5
    Employment
    □1 □2 □3 □4 □ 5
    Finance
    □1 □2 □3 □4 □ 5
    Health
    □1 □2 □3 □4 □ 5
    Tourism
    □1 □2 □3 □4 □ 5
    Culture
    □1 □2 □3 □4 □ 5

14. **Do you want to add any comment as to “how ready” do you think the Aruba Government is for language technologies in relation to eGovernment and Digital Transformation?**

Appendix B: Alternative Data Sources

Several classics translated in Papiamento by Aruban writer Jossy Mansur.

- Treasure Island - Robert Louis Stevenson (154 p.). *Source: University of Leiden*
- Robinson Crusoe – Daniel Defoe (138 p.). *Source: University of Leiden*
- Journey to the Center of the Earth – Jules Verne (142 p.). *Source: University of Leiden*
- The Count of Monte Cristo – Alexandre Dumas (160 p.). *Source: University of Leiden*
- Around the World in 80 Days – Jules Verne (144 p.). *Source: University of Leiden*
- 20,000 Leagues Under the Sea - Jules Verne (169 p.). *Source: University of Leiden*
- King Solomon's Mines – Hendry Haggard (186 p.). *Source: University of Leiden*

Jossy Mansur also wrote several dictionaries:

- Dictionary: Papiamento-Spanish (308 p.). *Source: University of Leiden*
- Dictionary: Papiamento-English (510 p.). *Source: University of Leiden*
- Dictionary 5 Languages: Papiamento-Dutch-English-Spanish-French (930 p.). *Source: University of Leiden*

Department of Education translated several books to Papiamento that were originally in English or Dutch to support multi-lingual education goals in Aruba.

- Catalogue ea.aw (complete)
- Example: Anne Frank's diary was one of the books that was translated Papiamento

Ministry of Education commissioned the translation the Gospel of Luke in Papiamento, in collaboration with Nederlands Bijbelgenootschap (NBG). The objective is to translate the complete bible by 2021.

The former fund (Fondo Desaroyo) – a joint Aruba/Netherlands development fund– sponsored the formalization of several books by professional linguists:

- Papiamento: prepositions and interjections. 2016
- Papiamento: orthography/vocabulary. 2009
- Papiamento: guidance on grammar/morphology. 2010